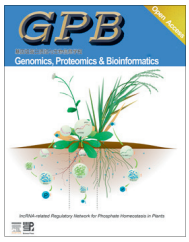
Genomics Proteomics Bioinformatics 13 (2015) 192–199

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



APPLICATION NOTE

BioCluster: Tool for Identification and Clustering of Enterobacteriaceae Based on Biochemical Data

Ahmed Abdullah [a], S.M. Sabbir Alam [b], Munawar Sultana [c], M. Anwar Hossain [*,d]

Department of Microbiology, University of Dhaka, Dhaka 1000, Bangladesh

Received 24 November 2014; revised 11 February 2015; accepted 10 March 2015
Available online 26 July 2015

Handled by Wuju Li

KEYWORDS

Bacterial identification;
Enterobacteriaceae;
Biochemical properties;
Clustering tool;
Identification tool;
Hierarchy algorithm

Abstract Presumptive identification of different **Enterobacteriaceae** species is routinely achieved based on **biochemical properties**. Traditional practice includes manual comparison of each biochemical property of the unknown sample with known reference samples and inference of its identity based on the maximum similarity pattern with the known samples. This process is labor-intensive, time-consuming, error-prone, and subjective. Therefore, automation of sorting and similarity in calculation would be advantageous. Here we present a MATLAB-based graphical user interface (GUI) tool named BioCluster. This tool was designed for automated clustering and identification of **Enterobacteriaceae** based on biochemical test results. In this tool, we used two types of algorithms, *i.e.*, traditional hierarchical clustering (HC) and the Improved Hierarchical Clustering (IHC), a modified algorithm that was developed specifically for the clustering and identification of **Enterobacteriaceae** species. IHC takes into account the variability in result of 1–47 biochemical tests within this **Enterobacteriaceae** family. This tool also provides different options to optimize the clustering in a user-friendly way. Using computer-generated synthetic data and some real data, we have demonstrated that BioCluster has high accuracy in clustering and identifying enterobacterial species based on biochemical test data. This tool can be freely downloaded at http://microbialgen.du.ac.bd/biocluster/.

Introduction

Enterobacteriaceae are a family of gram-negative, rod-shaped, facultative anaerobic bacteria, which are mainly recognized for their ability to cause intestinal diseases [1]. Enterobacteriaceae are responsible for a variety of human and animal illnesses, including urinary tract infections, gastroenteritis, meningitis, pneumonia, and septicemia [2,3]. Microbiological diagnosis for detecting the presence and type of Enterobacteriaceae from clinical samples is potentially important. Various biochemical

* Corresponding author.
E-mail: hossaina@du.ac.bd (Hossain MA).
[a] ORCID: 0000-0001-5550-9574.
[b] ORCID: 0000-0002-9148-3727.
[c] ORCID: 0000-0002-8563-3661.
[d] ORCID: 0000-0001-9777-0332.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.03.007

tests are traditionally used for presumptive identification and clustering of different enterobacterial species [4,5]. Biochemical tests such as indole production test, methyl red test, Voges-Proskauer test, citrate utilization etc. are usually performed [1,4–6]. Results of different tests are manually evaluated either as positive or negative for identification of a particular group of bacteria [2].

Manual check and comparison of biochemical test results are cumbersome and the results are sometimes hard to interpret, especially if the number of isolates is large. When there are a large volume of isolates, it becomes error-prone and difficult to reproduce, which is further confounded by the fact that the test result for a given species is not completely fixed: a given species may provide several combinations of biochemical test results [1].

By automating the analysis process of biochemical results, sorting (clustering), and identification of particular genera, the difficulties associated with manual sorting could be resolved. Here we propose a MATLAB-based tool, which was specifically designed for the clustering and identification of 128 species of Enterobacteriaceae from 30 genera based on the results of different biochemical tests (1–47 selected tests in Bergey's Manual of Systematic Bacteriology, [1]). We used two types of algorithms. One is the agglomerative hierarchical clustering algorithm (HC) and the other is a modified hierarchical clustering algorithm, which we termed as Improved Hierarchical Clustering algorithm (IHC). Agglomerative HC is a "bottom up" approach. Each observation starts in its own cluster, and pairs of clusters are merged as one to move up along the hierarchy [7,8]. Using BioCluster, HC can be applied directly to cluster Enterobacteriaceae isolates based on the biochemical properties. However, HC-based clustering may provide a misleading result due to the variability of the test results present within the same species. For example, closely-related *Escherichia* spp. can be sorted into different clusters, whereas different *Salmonella* spp. can be clustered with *Escherichia* spp. Therefore the algorithm was improved to take into account the variability of test results in Enterobacteriaceae by maximal utilization of relevant biochemical information for isolate clustering. We tested the accuracy of the new algorithm using computer-generated synthetic data and some real data, and it showed improved performance as compared to the naive HC algorithm.

BioCluster provides a user-friendly, easy method for the rapid clustering and identification of Enterobacteriaceae species based on biochemical properties. This tool is freely available for non-profit use at: http://microbialgen.du.ac.bd/biocluster/.

Methods

Algorithm

BioCluster uses HC as the clustering algorithm in two different ways. In one case, HC is directly applied to cluster the biochemical test results. However, biochemical test results are not numerical but are categorical (with binomial output as + or − for a given test). Hamming distance was thus chosen to measure the distance among different isolates, since it only considers the identity or non-identity of a test at a given position but not the actual numerical distance [9,10].

The clustering is further improved in IHC to provide a more refined output of biochemical test data. Species/isolates of a given genus show different levels of variability in their biochemical test results. For a given species, every test result may not be equally informative in clustering. For instance, the frequency for the *Edwardsiella hoshinae* isolates to produce a positive test result for indole production is about 50% [11]. So, the indole production test does not provide useful information for *E. hoshinae* identification/exclusion. As a result, the biochemical test results weight differentially when classifying a certain species.

Bergey's Manual of Systematic Bacteriology is a systematic catalog that contains information on the variability of the biochemical test results of a particular species [1]. Data tables for the frequency of positive biochemical test results for possible species of Enterobacteriaceae were taken from Bergey's Manual (Table S1) [2]. For IHC, the frequency table for positive results of biochemical tests (1–47 tests) is converted to conditional probability score matrixes for 128 Enterobacteriaceae species in 30 genera.

Naïve Bayesian model was used to find the probability of different instances of test results belonging to the set of 128 Enterobacteriaceae species [10,12,13]. It is assumed that the isolate belongs to one of the 128 members. If an isolate (*e.g.*, isolate 1) has the biochemical result T (T is a string of result, *e.g.*, $T = + - + + + \cdots -$), then probability score for species S_i is given by Bayesian probability as

$$P(S_i|T) = \frac{P(S_i)P(T|S_i)}{\sum_{j=1}^{n} P(S_j)P(T|S_j)} \quad (1)$$

Prior probability of being in one or other species is equal, which means:

$$P(S_1) = P(S_2) = P(S_3) = \ldots P(S_k) = \frac{1}{n} = 1/128$$

$$P(S_i|T) = \frac{\frac{1}{n}P(T|S_i)}{\frac{1}{n}\sum_{j=1}^{n} P(T|S_j)} = \frac{P(T|S_i)}{\sum_{j=1}^{n} P(T|S_j)} \quad (2)$$

There are t tests and test results are independent from each other according to the naivety assumption:

$$P(T|S_i) = \Pi_{j=1}^{t} Q_j \quad (3)$$

where Q_j stands for the probability of the Species S_i to show the same result for jth test as in T and n is the total number of species (128 in this case).

The choice of an appropriate distance metric is crucial for multidimensional clustering analysis. It is not always obvious what the distance metric means for a particular situation. In this study, we have chosen a distance metric, which we considered as one of the best possible solutions in the context, for the distance between the two isolates is defined as the probability that they belong to different species. It can be easily calculated from the conditional probability matrix. If C_i and C_j stands for conditional probability vector of ith and jth isolates, the distance is: